Appendix A

 Studies examining cycle shifts in women’s mate preferences have varied in numerous ways. Here, we describe typical study characteristics, noting common variations.

 *Sample Characteristics and Study Design Basics.* In most cases, samples are comprised of undergraduate students at a college or university. Less commonly, samples are comprised of community members or a combination of undergraduates and community members. In most studies to date, participants have been recruited in the US or UK (although we note exceptions in Table 2). Typically, study sessions take place in the lab. Somewhat less commonly, sessions take place in the field, by which we mean online or, very rarely, via a magazine survey with mail-in response. Either the same women complete sessions at high-fertility and low-fertility points in the cycle (a within-participants design), or different women complete sessions at high-fertility and low-fertility points in the cycle (a between-participants design).

 *Cycle Position Estimation Methods.* Researchers have used many methods to estimate and verify women’s position in the ovulatory cycle, and for the most part, there currently is not strong consensus on which of these methods or combinations of methods is most accurate and precise. Given that assessments of cycle position are crucial in determining the extent to which a study provides a precise measure of cycle shifts, we discuss these methods in detail here.

 Designing a study to examine cycle shifts entails unique challenges. Researchers cannot manipulate women’s hormones to mimic the naturally occurring changes that typically accompany (and drive) changes in women’s fertility across the ovulatory cycle. Therefore, researchers cannot conduct a true experiment to test whether shifts in women’s cycling hormones cause shifts in their mate preferences. Instead, researchers must capitalize on naturally occurring variation in women’s hormones across the cycle.

 In most previous studies that have examined cycle shifts in women’s mate preferences, researchers have used one of the following methods to schedule women’s sessions. In many cases, researchers have allowed participants to complete a single session at their convenience, without making any effort to have women participate on specific days of the cycle. Following their participation, researchers then used high- and low-fertility windows (usually defined in advance of data collection) to categorize women as high- or low-fertility or to exclude from analysis women who participated outside of both windows (e.g., Harris, 2010). Alternatively, researchers have sometimes assigned each woman a conception probability (fertility) value based on actuarial estimates (e.g., Wilcox, Dunson, Weinberg, Trussell, & Day Baird, 2001), thus retaining all women in the sample. Less commonly, researchers have allowed participants to complete their first session at their convenience but then scheduled subsequent sessions at some pre-designated interval (e.g., scheduling participants to complete their second session approximately two weeks after their first session; e.g., Johnston, Hagel, Franklin, Fink, & Grammer, 2001). In this case, researchers then use high- and low-fertility windows (usually defined in advance of data collection) to retrospectively categorize sessions already completed as high- or low-fertility or to exclude from analysis sessions that fell outside of both windows. Alternatively, some researchers have scheduled each participant to complete at least one session at a predicted high-fertility point and at least one session at a predicted low-fertility point based on scheduling windows defined in advance of data collection (e.g., Izbicki & Johnson, 2010, June; Garver-Apgar & Gangestad, *unpublished*).

 In order to schedule sessions in high- and low-fertility windows, retrospectively categorize women or sessions as high- and low-fertility, or retrospectively assign women conception probability values, researchers must estimate each participant’s position in the ovulatory cycle. Typically, researchers have used a “counting” method to estimate cycle position. The forward counting method typically entails identifying each participant's date of *last* menstrual onset, counting forward from that date—usually, by 14 days—to predict the date of her next ovulation, and then determining her current cycle position relative to that date (e.g., Little, Jones, Burt, & Perrett, 2007). A less common version of the forward counting method entails identifying each participant's date of last menstrual onset, counting forward from that date to her current day in the cycle, and then assigning her a conception probability value based on the number of days since her last menstrual onset (e.g., Gangestad & Thornhill, 1998). The reverse counting method entails predicting each participant’s date of *next* menstrual onset, counting back from that date—usually, by 14 or 15 days—to predict the date of her next ovulation, and then determining her current cycle position relative to that date (e.g., Puts, 2005).

 Large-scale studies have found that cycle length varies substantially both between and within women. However, studies have generally found that the follicular phase of the cycle (from last menstrual onset to ovulation) is more variable in length than is the luteal phase of the cycle (from ovulation to next menstrual onset; e.g., Baird et al., 1995; Fehring, Schneider, & Raviele, 2006). Because only the reverse counting method takes into account each participant’s unique cycle length, while holding constant the length of the luteal phase (and thus reducing the influence of variability in the length of the follicular phase on cycle position estimates), some researchers have asserted that the reverse counting method provides more accurate estimates of ovulatory cycle position than does the forward counting method. Nonetheless, regardless of which method is superior, it is noteworthy that the two counting methods often produce discrepant estimates of cycle position.

 Furthermore, both forward and reverse counting methods of cycle position estimation rely on a “benchmark” date of menstrual onset, which itself can introduce error into cycle position estimates. The forward counting method uses the date of last menstrual onset (preceding participation) as a benchmark. Typically, researchers obtain this date by asking participants to retrospectively recall it. The reverse counting method uses the date of next menstrual onset (following participation) as a benchmark. Researchers typically predict this date by asking participants to retrospectively recall their date of last menstrual onset and then adding either participants’ estimate of their average cycle length or a standard estimate of cycle length (e.g., 28 days) to that date.

 Notably, retrospectively recalled dates of menstrual onset appear to be error-prone. For example, in one study, researchers asked women to recall their date of last menstrual onset, which had also been prospectively collected within the past 30 days as part of a related study. Although 56% reported the correct date, 18% were off by 1 day, 7% were off by 2 days, and 19% were off by 3 or more days (Wegienka & Day Baird, 2005). We are not aware of any studies assessing the accuracy of women’s self-reports of their average cycle length, but it is possible that these self-reports are also error-prone. Given that women’s high-fertility window spans only about 6 days, and fertility decreases dramatically following this window, error such as that associated with participants’ retrospectively recalled dates of last menstrual onset and self-reported average cycle lengths could seriously compromise the precision with which researchers can estimate women’s cycle position and fertility.

 To safeguard against potential error introduced by women’s self-reported cycle information, some researchers have taken additional steps to ensure the accuracy of cycle position estimates. In some cases, researchers using the forwarding counting method have waited for participants to start a menstrual period (and report the start date to them) before scheduling their session(s). Similarly, some researchers using the reverse counting method have followed up with participants after they have completed the study (e.g., via regular phone calls or emails) to verify the date that they start their next menstrual period, which they then use to retrospectively categorize sessions already completed as high- and low-fertility and exclude observations collected outside of those windows (e.g., Roney et al., 2011).

 Occasionally, researchers have used rigorous methods to verify ovulation, rather than relying on counting methods alone to estimate cycle position. Although rare, researchers sometimes use the salivary ferning technique, which involves asking participants’ to lick or smear saliva on a slide in their predicted high-fertility window. Researchers examine the saliva under a microscope for ferning (crystallization) patterns that indicate impending ovulation (Fehring & Gaska, 1998). More commonly, researchers use the luteinizing hormone method, which involves having participants complete a series of urine tests to measure concentrations of luteinizing hormone in their predicted high-fertility window. Luteinizing hormone rises approximately 24-48 hours prior to ovulation (Testart & Frydman, 1982), and tests of LH in urine are highly accurate in verifying ovulation as detected by ultrasound (e.g., ClearBlueEasy tests are 97% accurate; Guermandi et al., 2001). Researchers using either of these ovulation verification methods typically either schedule participants’ sessions using the forward or reverse counting method and later exclude from analysis those participants who did not show evidence of impending ovulation in close proximity to their putative high-fertility session and/or who showed evidence of impending ovulation in close proximity to their putative low-fertility session (e.g., Gangestad, Thornhill, & Garver-Apgar, *unpublished*) or wait for participants to show evidence of impending ovulation prior to scheduling their sessions (e.g., Peters, Rhodes, & Simmons, 2009).

 In addition to the sources of variation already discussed, researchers using high- and low-fertility windows to schedule sessions or to categorize women or observations as high- and low-fertility must decide how to define these cycle phases. Given that there is not yet an established convention, the cycle shift literature is characterized by substantial variation across studies in the breadth and placement of high- and low-fertility windows. Women’s fertility begins to rise three to five days before ovulation. Fertility is at a maximum on the day before and the day of ovulation and decreases dramatically the day after ovulation (Wilcox, Dunson, Weinberg, Trussell, & Day Baird, 2001). Women are effectively nonfertile for the remainder of the cycle. Thus, how a researcher defines the low-fertility window might not be very influential in determining whether a study provides a precise measure of cycle shifts. As long as the low-fertility window is not so badly misplaced or overly broad that it includes high-fertility days, it is likely to capture women on true low-fertility days. In contrast, given that women are effectively fertile on only a handful of days in the ovulatory cycle, how a researcher defines the high-fertility window is crucial in determining whether a study provides a strong test—or *any* test—of cycle shifts. A misplaced or overly broad high-fertility window can result in women completing putative "high-fertility" sessions on true low-fertility days of the cycle. This is analogous to failing to expose participants in a test condition to the manipulation and, therefore, comparing controls to controls.

 To attempt to assess the precision of the various high- and low-fertility windows that have been used by the studies included in this meta-analysis, we estimated the average conception probability of the high- and low-fertility window for each study using actuarial estimates of the probability of conceiving from a single act of unprotected sexual intercourse on a given day of the ovulatory cycle. For studies that verified ovulation using luteinizing hormone tests, we contacted study authors to obtain the data needed to estimate each woman’s conception probability at her high-fertility session based on her proximity to ovulation (Weinberg and Wilcox, 1995). In order to convert days from the luteinizing hormone surge to days from ovulation, we added one day (Testart & Frydman, 1982). For example, if a woman completed her high-fertility session on the same day she showed a luteinizing hormone surge, we converted this to one day before ovulation and assigned her a conception probability estimate of 0.31 (see Weinberg & Wilcox, 1995). We then averaged across all women in that sample to obtain an estimate of the average conception probability of the high-fertility window. We estimated the average conception probability of the low-fertility window based on the forward cycle days included in that window (Wilcox, Dunson, Weinberg, Trussell, & Day Baird, 2001). Likewise, for studies that used a forward or reverse counting method (without verifying ovulation), we estimated the average conception probability of the high-fertility and low-fertility windows based on the forward cycle days included in those windows (Wilcox et al., 2001). In order to convert reverse cycle days (days until next menstrual onset) to forward cycle days (days since last menstrual onset), we assumed a cycle length of 28 days. For example, if a study defined the high-fertility window as including reverse cycle days 14-21, we first converted that to forward cycle days 7-14 and then took the average of the conception probability values (Wilcox et al., 2001) associated with those eight days: (.018 + .032 + .050 + .069 + .085 + .094 + .093 + .085) / 8 = 0.068. For this example, we would interpret the average conception probability estimate of 0.068 as indicating that women in the study’s high-fertility window had an estimated 7% probability of conceiving from a single act of unprotected sexual intercourse.

 Although the technique just described was useful to the extent that it enabled us to compare the precision of high-fertility and low-fertility windows across studies using very different cycle position estimation methods, our estimates of the average conception probability of high-fertility and low-fertility windows are based on several assumptions that are unlikely to be fully met and should therefore be regarded with due caution. First, our method of converting days from the LH surge to days from ovulation assumes that one day passes between the LH surge and ovulation, which is not the case for all women and all cycles. Second, our method of converting reverse cycle days to forward cycle days assumes an average cycle length of 28 days, which is likewise not the case for all women and all cycles. Lastly, our method of computing the average conception probability of a given high- or low-fertility window assumes that an equal number of women participated on each day of that window, when in fact that is unlikely to be the case, perhaps especially in small samples. Therefore, our average conception probability estimates are likely to include some error, and the methods we present here are in need of further honing and validation.

 *Determining the Amount of a Characteristic in Male Stimuli.* Researchers have used several methods to vary or measure the amount of a characteristic possessed by male stimuli. In some studies, researchers have selected stimuli that naturally vary on the characteristic of interest and then either directly measured the characteristic of interest (e.g., used calipers to measure facial masculinity), obtained reliable codes of the characteristic (e.g., trained research assistants to reliably code specific behaviors in participant videos), collected ratings of the characteristic from participants in the cycle study (e.g., asked participants in the cycle study to rate how masculine the faces looked), or collected ratings of the characteristic from a separate sample of participants. In other studies, researchers have directly manipulated stimuli to vary the characteristic of interest (e.g., morphed faces to increase or decrease symmetry) or allowed participants to directly manipulate a characteristic in stimuli.

 *Mate Preference Measures.* Researchers have assessed women’s mate preferences using numerous methods. Typically, researchers have used measures of *revealed preferences* to assess women’s attraction to a given characteristic. By our definition (see Wood & Brumbaugh, 2009, for a similar but somewhat narrower definition), this entails either a) asking women to rate or choose (e.g., in a two-option or multiple-option (3 or more) forced choice) among men or male stimuli known to vary on a characteristic and then examining the association between women’s reported attraction to the stimuli and the amount of the characteristic possessed by those stimuli or b) asking women to directly manipulate a characteristic until they have achieved what they perceive to be the most attractive version of a male stimulus (e.g., a facial photograph) and then examining the amount of the characteristic deemed most attractive. Less commonly, researchers have used *stated preference* measures to assess women’s attraction to a given characteristic. By our definition, this entails asking women to explicitly rate the importance or desirability of a characteristic in a prospective partner. Regardless of which of these measures researchers have used, they have sometimes also asked women to make their evaluations in reference to a specific relationship context (discussed in more detail in the Methods section).